The Grid Observatory 3.0 - Towards reproducible research and open collaborations using semantic technologies

Cécile Germain, Julien Nauroy, Karima Rafes

EGI Community Forum 2014 - Helsinki

Abstract ID : 69

The Grid Observatory 3.0 - Towards reproducible research and open collaborations using semantic technologies

Summary :

The Grid Observatory 3.0 evolves the Grid Observatory (G.O.) and Green Computing Observatory (G.C.O.) along the Open Linked Data and reproducible research concepts.
The first objective is to make analysis easier and more productive, by addressing the technical heterogeneity of the data (EGI services logs), and the wide range of potential usage. Semantic web technologies address these by (i) creating an OWL ontology of the EGI software architecture, (ii) converting the traces from selected services of the grid into an ontology compatible RDF format and (iii) organizing them in SPARQL-enabled triple stores. These technologies expedite and make transparent the personalized integration of multiple, independent sources, required for analysing the behaviour of the EGI grid, as well as long-term sustainability of the GO and GCO repositories. Moreover, the scientist's activity can be exploited to refine the ontology in a collective knowledge building process
The second objective is to encourage reproducible science by providing ways to repeat in silico experiments based on the GO data and stored queries over data and processing algorithms. A catalogue of customizable queries will be provided to show examples of queries and processing over the published data. Cloud-based hosting and processing capabilities will be offered to scientists to store and share their processes and algorithms through a collaborative platform in order to encourage open collaborations.

Description :

Based on our previous experience with the GO and GCO, we are currently building an ontology of the EGI grid, with a focus on the time series acquired from different data sources (BDII, L, GridFTP, WMS etc). This ontology, represented in an OWL format, allows the interpretation of data from different sources by creating bridges between them. As an example, the queues defined in the EGI-wide BDII have their counterpart in each WMS, though the association is not straightforward. By linking the EGI and WMS queues, it becomes possible to follow individual jobs from the time they are submitted to EGI to the moment they are executed on a particular node. After conversion, the data will be hosted on a publicly-available cluster of triple stores, serving data in RDF format through SPARQL queries. We expect to publish a few billion RDF triples resulting from the conversion of our history of traces.
The second step is to help scientists query over these data and run experiments. We have started a collaborative Wiki with enhanced functionalities to allow for querying data through SPARQL queries. The goal is to help scientists create queries collaboratively and get help when needed. We will also provide a catalogue of queries to help users getting started and provide them with solutions to common demands. One of the intermediate steps and a test for our system would be to reproduce an experiment previously published, involving the processing of both BDII and WMS traces.
The last step will be to provide ready-to use workflows to automatically compute the results of an experiment, given the SPARQL query to execute in order to construct the dataset and the algorithm to apply to the resulting data. We envision the automatic provisioning of computing resources from a cluster such as Stratuslab to host the dataset resulting from the

query, run the algorithm and return the results to the portal.

Conclusion :

The G.O. 3.0 positions itself in the domain of Semantic Web technologies by publishing data in Linked Data format. Hosting data in publicly-available triple stores opens up new possibilities to interconnect distinct datasets previously represented in incompatible formats. Multiple datasets can be queried at the same time through the use of federated SPARQL queries, allowing for an easier persistence of large datasets. Providing users with workflows to execute algorithms over these datasets allows for reproducible research as results from an experiment could be recalculated over the original data. It further allows studying the behaviour of an algorithm over a new dataset simply by changing the query building it, or comparing two different algorithms by providing them with the exact same dataset. Sharing the queries and results on such workflow-enabled Wiki will push this concept further by encouraging scientists have their experiments verified and confirmed by peers.

URL :

http://grid-observatory.org

Primary authors : GERMAIN-RENAUD, Cecile (CNRS) ; NAUROY, Julien (CNRS) ; RAFES, Karima (Inria)

Co-authors :

Presenter : NAUROY, Julien (CNRS)

Track classification : Data and knowledge preservation and curation (Track Leaders: J. Shiers, A. Fresa)

Contribution type : Sessions contributions

Submitted by : NAUROY, Julien

Submitted on Saturday 15 February 2014

Last modified on : Saturday 15 February 2014

Comments :

Not sure if our proposed talk should enter in the "session" or "sessions contributions" presentation type. Feel free to change it if the one I chose doesn't match.